Learning to Identify Local Flora with Human Feedback

Stefan Lee and David Crandall
School of Informatics and Computing
Indiana University
{steflee,djcran}@indiana.edu

1. Introduction

Plants and animals frequently appear in consumer images but are often incidental background objects whose specific fine-grained details cannot be seen. For instance, consider the photo on the left side of Figure 1 – what species of tree is highlighted in red? The answer to this question could provide useful information about the photo for a range of applications. Photo organization software could automatically tag images with species names of flora or fauna to support content-based retrieval [10]. Detecting and identifying species could help to infer a geo-tag for an image [4], especially for rural photos that lack other geo-informative evidence, since many species of plants and animals occur only in certain regions of the world [3]. On the other hand, when images are already geo-tagged, recognizing species could support citizen science applications that use consumer photos to track the distribution of natural phenomena [8].

But flora identification is a very difficult problem, both for computers and for humans that are not domain experts. (Did you correctly identify the tree in Figure 1 as a Chilean Wine Palm, or *Jubaea chilensis*, which is endemic to central Chile?) While recent work has considered automated techniques for fine-grained classification, including classifying among species of birds [9] and leaves [6], these papers typically study images in which the objects of interest are large and have distinctive local features (like shapes of individual leaves) that are readily visible. Other recent work has built hybrid human-computer recognition systems, using mid-level visual attributes (image features that are both visually distinctive and semantically-meaningful) as the "language" to allow humans and computer vision algorithms to collaborate on recognition tasks [1, 2]. These techniques work well in domains where clean common-language visual attributes exist, as in bird recognition with attributes like "yellow beak" and "white belly." But these techniques are hard to apply with non-expert users who lack the vocabulary for describing properties of objects, especially when individual properties of the object are not visible and recognition must rely on the overall "look" of the object (as in Figure 1). This challenge is confounded by the fact that specimens of the same species differ widely in appearance, e.g. in plants due to factors such as age, climate, disease, pruning, etc.

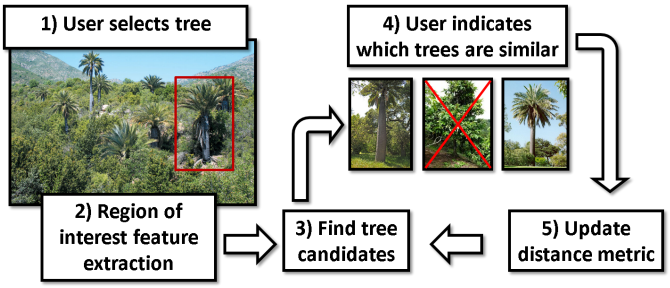

Figure 1: Diagram of our proposed human-in-the-loop system. We take a hand-marked target tree, extract features and find visually similar trees in a library of annotated images, ask the user for feedback on which of these candidates are similar, update the distance metric, find new candidates, and iterate until convergence to a tree label.

In this ongoing work, we are developing a method that involves a user in the loop to aid in the fine-grained recognition of a diverse set of tree species. Instead of asking users to provide attributes of trees, we instead ask them to judge the similarity between pairs of tree images, and then use this to learn the parameters of a discriminative distance metric for use with k-nearest neighbors. Over time, the discriminative distance function becomes a better approximation to the human's judgment of visual similarity. We present baselines and results of our human-guided approach on a collection of 20 tree species from five geographic locations.

2. Methodology

Our approach has an offline and online training phase. We assume that we have a labeled training set of images that are cropped tightly around single tree exemplars. In the offline phase, we extract global features from each cropped training image and use the known labels in the training set to learn the parameters of a distance metric. We use the regularized online distance metric learning algorithm of Jin

Figure 2: Human interaction GUI. Given a target image (top), the user is shown candidate matches under the current distance metric (bottom), and is asked to indicate which images appear to be true matches.

et al [5] for both offline and online learning. Given a pair of exemplars with known class labels, the approach minimizes a regularized loss function based on the squared Mahalanobis distance as a function of the covariance matrix. We adapt this approach to a batch method by using it as a sub-routine in a pocket-algorithm fashion. We iterate over all pairs in the training set and evaluate the total error. As in the pocket-algorithm we retain the best solution so far until many iterations without improvement are observed.

In the online phase, human interaction is used to improve the distance metric and recognition results. The human user selects an image region corresponding to an unknown tree of interest. We compute global features like GIST [7] from that region, and find the k most similar images in our training set to present to the user. The user then indicates whether each of these tree images appears to be of the same species as the target image (Figure 2). Using these positive (objects are similar) and negative (objects are dissimilar) responses, the algorithm updates the distance metric using [5] and presents the user with the new k-nearest neighbors under this updated distance metric. This cycle repeats until the user thinks most candidates are similar to the target image, in which case the system suggests the majority label.

3. Evaluation

For a preliminary evaluation, we chose four indigenous tree species from each of a diverse set of five countries (Philippines, Chile, Jordan, India, and Taiwan) to create a 20-way classification problem. We collected a dataset of 269 images (from Flickr and the web) distributed approximately evenly over the tree classes. We withheld about 10% of these images as a test set, and cropped the remaining images around the tree exemplars to produce our training set.

We first evaluated a fully-automatic recognition approach. Using only GIST features and a nearest-neighbor classifier under Euclidean distance, we achieved a classification accuracy of 15.8%, relative to a 5% majority-class baseline. After learning a new distance metric using only offline training, the fully-automatic accuracy increased to 26.3%. We then evaluated the human-in-the-loop technique using a simple GUI and a non-expert human user. The user was asked to interact with the system, iteratively selecting visually-similar images (which the system was using to update the distance metric) until he or she believed that most of the candidates were visually similar to the target image, and then the system assigned that label. The user attained an accuracy of 36.8%, or over seven times baseline, on this challenging fine-grained categorization task.

4. Conclusion

Our preliminary results demonstrate the potential of a human-in-the-loop approach to solve a challenging tree recognition problem that would be difficult or impossible for computers or humans to solve individually. This is ongoing work and we are continuing to explore a variety of directions, including using more sophisticated visual features, injecting diversity into the sets of candidates, and studying other fine-grained classification tasks.

Acknowledgments. Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] S. Branson, G. Van Horn, C. Wah, P. Perona, and S. Belongie. The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization. *IJCV*, 2014. 1

[2] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012. 1

[3] B. Groombridge and M. Jenkins. *World Atlas of Biodiversity*. UNEP-WCMC, 2002. 1

[4] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *CVPR*, 2008. 1

[5] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *NIPS*, 2009. 2

[6] N. Kumar, P. Belhumeur, A. Biswas, D. Jacobs, W. J. Kress, I. Lopez, and J. Soare. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*, 2012. 1

[7] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res.*, 155, 2006. 2

[8] J. Wang, M. Korayem, and D. Crandall. Observing the natural world with Flickr. In *ICCV Workshops*, 2013. 1

[9] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, CalTech, 2010. 1

[10] D. Zhang, M. Islam, and G. Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1), 2012. 1